CodeT5+: Open Code Large Language Models for Code Understanding and Generation

Year: 2023 | *Citations: 597* | *Authors: Yue Wang, Hung Le, Akhilesh Deepak Gotmare*

Abstract

Large language models (LLMs) pretrained on vast source code have achieved prominent progress in code intelligence. However, existing code LLMs have two main limitations in terms of architecture and pretraining tasks. First, they often adopt a specific architecture (encoder-only or decoder-only) or rely on a unified encoder-decoder network for different downstream tasks. The former paradigm is limited by inflexibility in applications while in the latter, the model is treated as a single system for all tasks, leading to suboptimal performance on a subset of tasks. Secondly, they often employ a limited set of pretraining objectives which might not be relevant to some downstream tasks and hence result in substantial performance degrade. To address these limitations, we propose ``CodeT5+'', a family of encoder-decoder LLMs for code in which component modules can be flexibly combined to suit a wide range of downstream code tasks. Such flexibility is enabled by our proposed mixture of pretraining objectives to mitigate the pretrain-finetune discrepancy. These objectives cover span denoising, contrastive learning, text-code matching, and causal LM pretraining tasks, on both unimodal and bimodal multilingual code corpora. Furthermore, we propose to initialize CodeT5+ with frozen off-the-shelf LLMs without training from scratch to efficiently scale up our models, and explore instruction-tuning to align with natural language instructions. We extensively evaluate CodeT5+ on over 20 code-related benchmarks in different settings, including zero-shot, finetuning, and instruction-tuning. We observe state-of-the-art (SoTA) model performance on various code-related tasks, such as code generation and completion, math programming, and text-to-code retrieval tasks. Particularly, our instruction-tuned CodeT5+ 16B achieves new SoTA results on HumanEval code generation task against other open code LLMs.